

Analysis of

Ride Cancellation

Data:Taxi-cancellation-case.csv

# **Executive Summary**

# This business report aims to assist Yourcabs.com, an India-based taxi company, in predicting driver trip cancellations and offering recommendations to address this issue. Driver cancellations and no-shows cause customer inconvenience and dissatisfaction, impacting the company's revenue and reputation. Our team used three modeling techniques to identify the major factors causing cancellations: logistic regression, classification trees, and random forests. We used the taxi-cancellation-case dataset and divided it into training and validation datasets. The models' accuracy, error, gain, and lift were assessed to determine their effectiveness in predicting car cancellations. During our analysis, we identified various relationships between factors and cancellations and the importance of these factors in the model. Our team recommends implementing targeted strategies based on insights gained from the models to improve customer satisfaction and reduce cancellations. By implementing these recommendations, Yourcabs.com can develop more accurate and reliable predictive models for car cancellations, enabling the company to take preventive measures and enhance customer satisfaction.

**Data preparation**

# **Dataset description**

# The original taxi-cancellation-case dataset consists of 10,000 rows and 20 variables. Among 10,000 bookings, there were 743 canceled bookings and 9257 not canceled bookings. For this dataset, some variables like to\_city\_id have 9661 missing values; other variables like package\_id have 8248 missing values, from\_city\_id has 6294 missing values, and to\_data has 4178 missing values, and some other variables contain missing values.

# 

**Data cleaning**

Data cleaning was performed to ensure the data set used for model building was accurate and reliable. Our team removed and consolidated variables with many missing values to improve data quality. The variables from\_lat, from\_long, to\_lat, and to\_long were combined to create the trip length variable while waiting time was calculated by subtracting booking\_created time from from\_data. A new variable, traditional\_booking, was created when both online\_booking and mobile\_site\_booking were equal to zero. The date and time variables were separated to find the week and month of the day.

Furthermore, our team explored the relationship between vehicle\_model\_id and the number of cancellations. As indicated in the left bar chart in Exhibit 1, most bookings were from vehicle model id 12, with a higher cancellation rate as the data increased. After assessing the importance of the input variables, we concluded that vehicle\_model\_id was not a significant factor in predicting cancellations and therefore removed it from the new dataset.

The new dataset consisted of 11 variables and 10,000 bookings, grouped into three categories: travel types, time, way of booking, and distance. This data-cleaning process ensured that our models were built on high-quality data, enhancing the accuracy and reliability of our predictions.

**Data partitioning**

After completing the data cleaning process, the team divided the dataset into training and validation sets. The training set accounted for 60% of the data, and the remaining 40% was allocated to the validation set. This partitioning approach was implemented to avoid overfitting the data.

The dataset was modified accordingly to build our three classification models, namely logistic regression, classification tree, and random forest. Specifically, we normalized the data using the range method for the logistic regression model. We oversampled the train data for the classification tree model by selecting 90% of the canceled bookings from the dataset.

**Descriptive Analysis**

**Barplot for Car\_Cancellation & travel type**

Exhibit 2 shows the relationship between car cancellation and travel types. The values “0” and “1” on the x-axis represent non-cancelation and cancellation, respectively. The right graph represents the number of cancellations based on different travel types. Among these, the travel type “point to point” has the highest number of car cancellations. As a result, the team assumes that point-to-point might be a factor that could influence car cancellation.

**Plots for Car\_Cancellation & time (months & weeks)**

These two graphs (Exhibit 3) show the trend of car cancellation based on the period. The left one shows the trends of cancellations from each month as it is shown that May, October, and November have the highest number of cancellations. On the other hand, the right graph shows the car cancellations from the view of weekdays (Mondays - Sundays). Among these, Fridays and Sundays have the highest number of cancellations. From this information, the team could know that the number of car cancellations increases during a certain period.

**Predictive Analysis**

**Model 1: Logistic Regression Model**

Car\_Cancellation = -0.86950\*Long Distance - 0.86950\*PointToPoint +

0.16819\*month + 1.16568\*online\_booking + 1.16123\*mobile\_site\_booking - 4.67131

One of the three classification models we built is the logistic regression model because it is easy to implement and interpret yet efficient in training and can be updated easily with new data using stochastic gradient descent. We used both-way stepwise selection technique that filtered 5 significant variables.

Generated equation in terms of these important predictors shows that long distance and point to point has a negative relationship with the car cancellation. In the other words, as the distance between two points increases, the likelihood of drivers canceling an order decreases. It could be interpreted in the context of rideshare or delivery services. One possible reason for this relationship could be that longer distances typically result in higher fares or fees, which can be more appealing to drivers due to increased earnings. As a result, they might be less inclined to cancel orders involving longer distances. However, it is important to note that this relationship may not hold true in all cases, and various factors like traffic, weather, and individual driver preferences can impact this relationship.

However, month,online booking, and mobile site booking are positively impacting the car cancellation. It seems that you are suggesting that certain factors, such as the month, online booking, and mobile-site booking, have a positive impact on car cancellations. These factors can indeed contribute to changes in cancellation rates for rideshare or delivery services. Let's examine each factor in more detail:

* Month: Cancellation rates may vary across different months due to seasonal factors. For example, during holiday seasons, demand for rides or deliveries could increase, leading to higher cancellation rates as drivers become more selective about which orders they accept. Conversely, during off-peak months, drivers might be less selective, and cancellation rates could decrease.
* Online booking: Online booking systems have made it easier for customers to book rides or deliveries, often at the click of a button. This convenience can lead to an increase in bookings, which might result in higher cancellation rates if drivers are faced with a larger pool of orders to choose from. Additionally, the ease of online booking may encourage some customers to make impulsive decisions, leading to more cancellations if they change their minds later.
* Mobile-site booking: Similar to online booking, mobile-site booking offers customers the convenience of booking rides or deliveries on the go. With more people using smartphones for various tasks, including transportation and delivery services, the increased accessibility could lead to higher booking volumes and, subsequently, higher cancellation rates.

The logit model has an accuracy rate of 92% for the training dataset and 93% for the validation dataset. This suggests that the model is generally effective in predicting car cancellations. However, the model appears to perform differently when it comes to sensitivity and specificity, which indicates that these factors may not always lead to increased cancellation rates, as the relationship between these variables and cancellations can be influenced by other factors, such as market conditions, company policies, and driver preferences. Sensitivity (true positive rate): with a sensitivity rate of 1 for both datasets, the model is perfectly predicting all actual non-cancelled cases as non-cancelled(please note how the confusion matrix calculated specificity in this case, as shown in Exhibit 1). This is a positive indication that the model is accurately capturing the factors contributing to non-cancelled cases. Specificity (true negative rate): with a specificity rate of 0 for both datasets, the model is not accurately predicting any of the actual canceled cases as canceled. This is a significant limitation of the model, as it means that it is not effectively identifying the factors leading to car cancellations.

Decile-wise Lift is the ratio of the gain percentage relative to the mean random result. Decile chart(Exhibit 4) shows that 73%/75% of the events are in 40% of the data. Thus, by focusing on 40% of the data, we can be more efficient.

**Model 2: Classification Tree**

After the logistic regression model, our team created a classification tree as they are generally easier to interpret than logistic models and have a higher rate of predicting the correct cancellation status. Then, knowing the dataset imbalance, we oversampled the training data by selecting 90% of the canceled bookings and validated the model using the original validation set.

To prevent overfitting and improve generalization, we created a deep tree and printed the tree attribute complexity parameter tables to show the results recorded during the tree-growing process. Eventually, we found that the tree has the smallest cross-validated error (xerror) when cp equals 0.00403769, as shown in Exhibit 5.

The first decision is based on the variable "traditional\_booking," indicating when both online\_booking and mobile\_site\_booking are equal to zero. The next decision is based on features such as "month," "waiting," and so on. Based on the insights gained from the classification tree model, Yourcabs.com can identify the most critical factors affecting car cancellations and implement targeted strategies to address these issues, ultimately improving customer satisfaction and reducing cancellations.

The classification tree model achieved an overall accuracy of 89.38% on the training set and 92.95% on the validation set. The sensitivity for cancellation cases was high on both the training and validation sets, with values of 97.70% and 97.56%, respectively. However, the specificity for normal cases was low on both sets, with values of 30.55% and 30.29%, respectively, which suggests that the model struggles to identify normal cases accurately.

On the training set, the model achieved a lift of 527 when targeting the top 6% of the cases, which means the model is 5.27 times better than random selection. On the validation set, the model achieved a lift of 696 when targeting the top 4% of the cases, which means the model is 6.96 times better than random selection. The lift indexes are relatively consistent between the training and validation sets, suggesting that the model may not overfit the training data.

**Model 3: Random Forest**

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to make a final prediction. The main advantage of the random forest model is that it provides more accurate predictions than a single decision tree. However, there may be an overfitting problem if the dataset is small. Our team has decided to build a random forest model to predict car cancellations based on other explanatory variables. We will test the model's performance using a confusion matrix, gain chart, and lift chart to check its generalization ability.

The model achieved high accuracy on both the training (93.3%) and validation (93.4%) sets, with high sensitivity for cancellation cases (99.96% and 99.68%, respectively). However, the model's low specificity for normal cases(14.71% for training and 8.39% for validation) suggests that it struggles to accurately identify normal cases. This, along with the decrease in both sensitivity and specificity on the validation set, indicates a potential overfitting issue. The model may perform well on the training data but struggle with new data, which can negatively impact its real-world effectiveness.

To investigate further, we used gain charts and lift charts to evaluate the model's performance on the validation data and check whether the model's performance is consistent across both the training and validation data sets. The gain chart analysis on the training dataset shows that the model achieved a lift of 854 when targeting the top 10% of the cases, which means that the model is 8.54 times better than random selection. The gain chart analysis on the validation dataset shows that the model achieved a lift of 453 when targeting the top 10% of the population, which means that the model is 4.53 times better than random selection. The discrepancy in lift index performance when targeting the top 10% of the cases between the training and validation datasets suggests that the model may be overfitting to the training data. While the model can effectively identify cancellation in the training data, its performance is poorer in the validation data. Further analysis is needed to confirm the cause of this difference and to improve the model's generalization ability.

**Models Performance Comparison**

|  |  |  |  |
| --- | --- | --- | --- |
| **Evaluation Metrics** | **Logistic Regression** | **Classification Tree** | **Random Forest** |
| Accuracy Rate (Training data) | 0.9218 | 0.8938 | 0.933 |
| Accuracy Rate (Validation data） | 0.9315 | 0.9295 | 0.9342 |
| Sensitivity(Training data) | 1 | 0.977 | 0.9996 |
| Sensitivity (Validation data) | 1 | 0.9756 | 0.9968 |
| Specificity (Training data) | 0 | 0.3055 | 0.1471 |
| Specificity (Validation data) | 0 | 0.3029 | 0.08394 |

The table below summarizes the accuracy rates, sensitivity, and specificity of three models – logistic regression, classification tree, and random forest on both training and validation datasets– in predicting car cancellations for Yourcabs.com. Based on the table, all three models generalize well to new data because of their consistent accuracy rate, sensitivity, and specificity.

The logistic regression model performed consistently on both training and validation datasets. However, it is worth noticing that with a specificity rate of 0 for both training and validation datasets, the model is not accurately predicting any of the actual canceled cases as canceled(please note how the confusion matrix calculated specificity in this case, as shown in Exhibit 1). Its failure to accurately identify true positive cases is a critical weakness of the model since its effectiveness hinges on its ability to pinpoint the underlying factors contributing to car cancellations.

On the other hand, although the random forest model achieved the highest accuracy rates on both datasets, it also exhibits signs of overfitting, as seen in its inconsistent performance between the top 10% cases in the gain chart (see Exhibit 9). This limitation could impact the model's ability to predict car cancellations on future data accurately. Therefore, it might be beneficial to further tune the random forest model by adjusting its hyperparameters, such as tree numbers, to reduce overfitting while maintaining its high accuracy rate.

Among these three models, the classification tree model performs better than the regression model and the random forest as its consistent evaluation metrics and its consistent slope for gain chart. Moreover, it has an acceptable accuracy rate and less probability of overfitting. On the other hand, the model is easier to understand and interpret as it is a white box model that shows all the split processes.

**Recommendation**

Based on our extensive analysis using various models and calculations, our team has identified four key recommendations for YourCabs. These recommendations aim to address the issue of high cancellation rates, which is a significant challenge for the company.

* User Notification System: To address the issue of high cancellation rates, one potential solution is to implement a user notification system. This would inform customers of the likelihood of their bookings being canceled, helping to manage their expectations and enabling them to make informed decisions.
* Booking Reconfirmation System: In order to reduce the number of last-minute cancellations, we propose implementing a booking reconfirmation system. This would give users the option to confirm or cancel their bookings in cases where there is a high likelihood of cancellation, resulting in a more reliable service for customers.
* Service Fee Reduction: One approach to incentivizing drivers to accept bookings rather than canceling them is to reduce the service fees charged during peak cancellation periods. This would not only encourage more consistent customer service but also attract more potential consumers and strengthen the company's reputation.
* Transparent Policies: To improve the company's policies around trip cancellations, we recommend conducting a review and revision process. This could include updating penalties for drivers who frequently cancel trips and providing clearer guidelines for drivers to follow. By implementing more transparent policies, YourCabs can deter drivers from unnecessary cancellations and ensure a more reliable service for its customers.

By implementing these recommendations, YourCabs can significantly enhance its operations and address the challenge of high cancellation rates. In addition, this would result in a more positive customer experience, leading to increased customer loyalty and a stronger brand reputation. Ultimately, these modifications would help YourCabs build a more successful and sustainable business.

**Appendix:**

Exhibit 1: Barplot for Car\_Cancellation & vehicle\_model\_ids

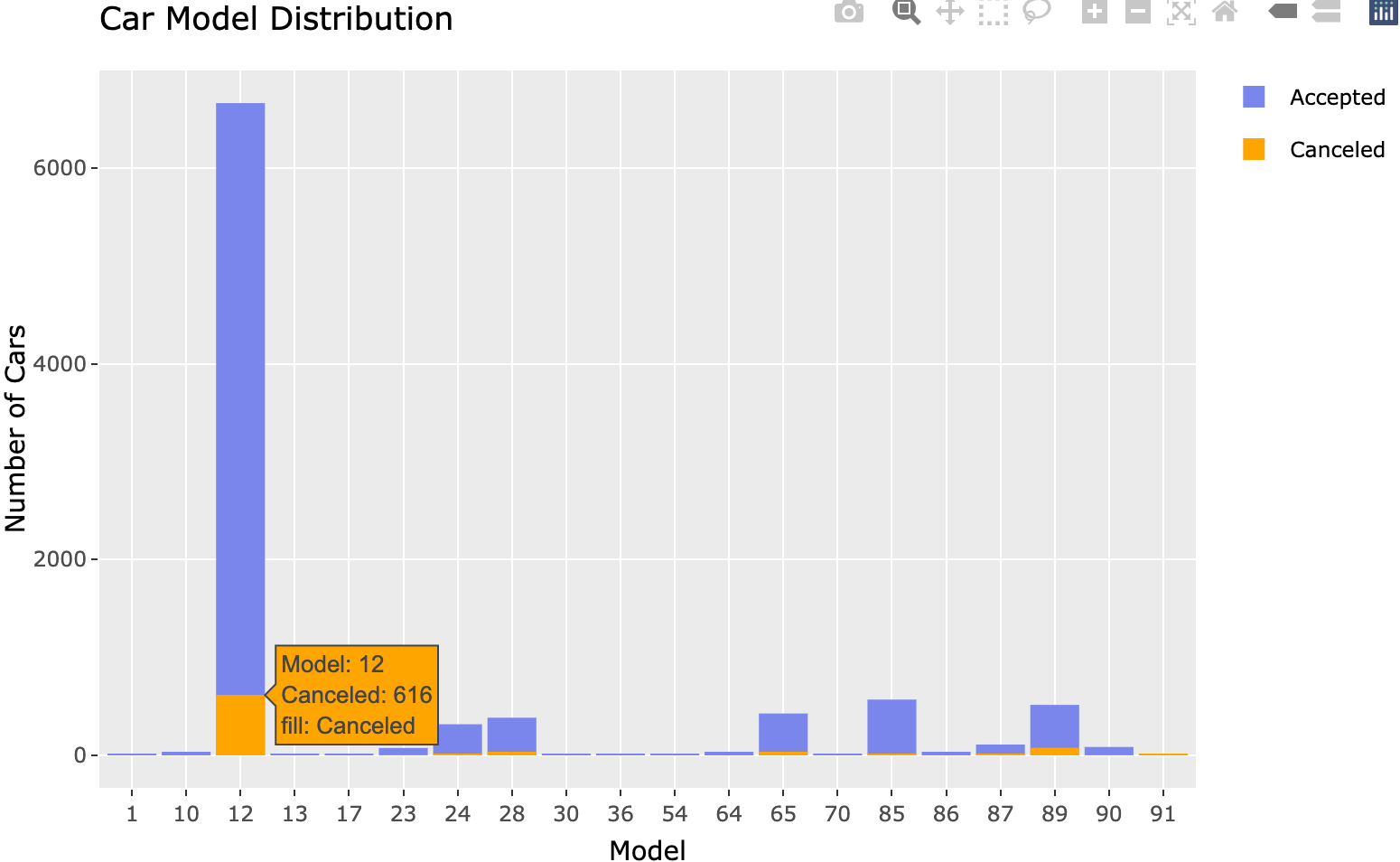


Exhibit 2: Barplot for Car\_Cancellation & travel type

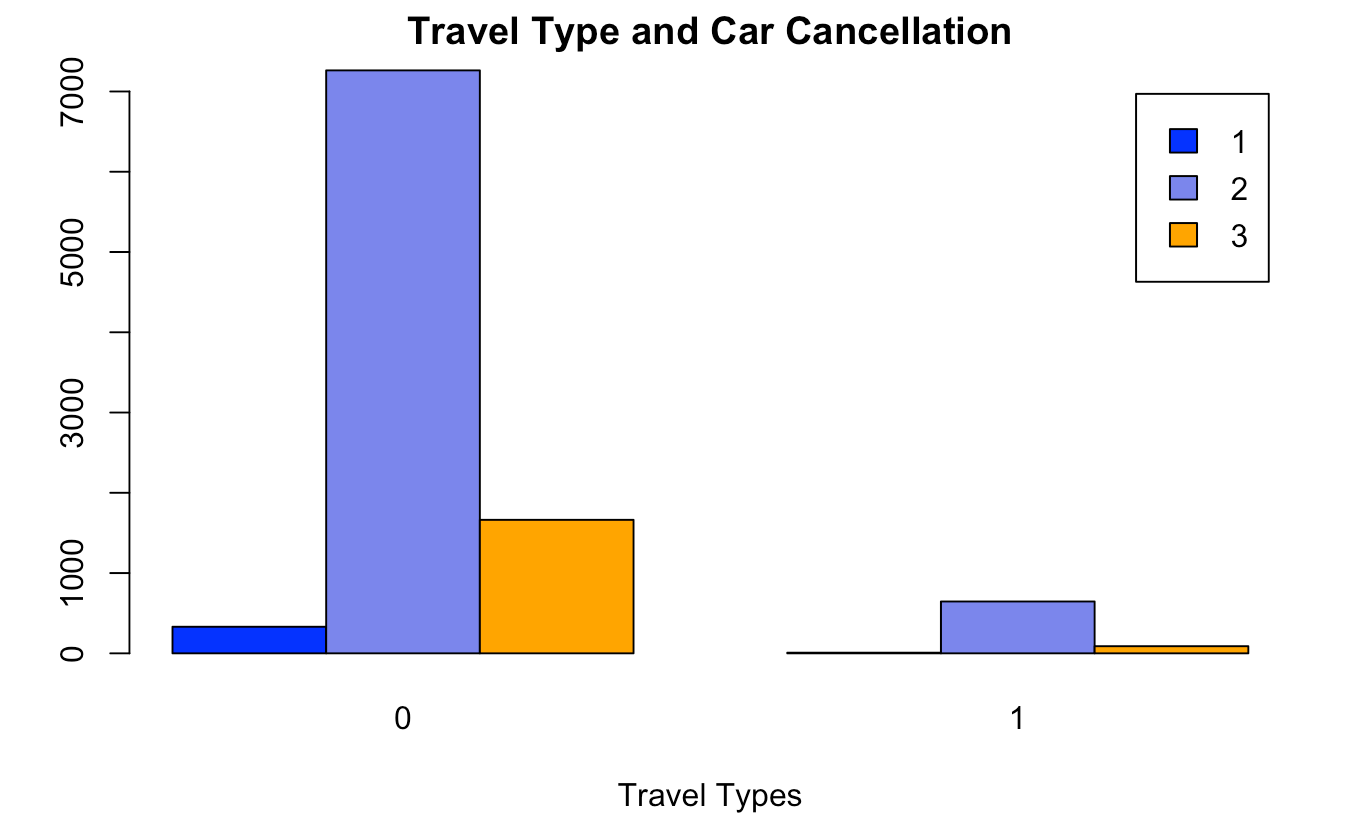
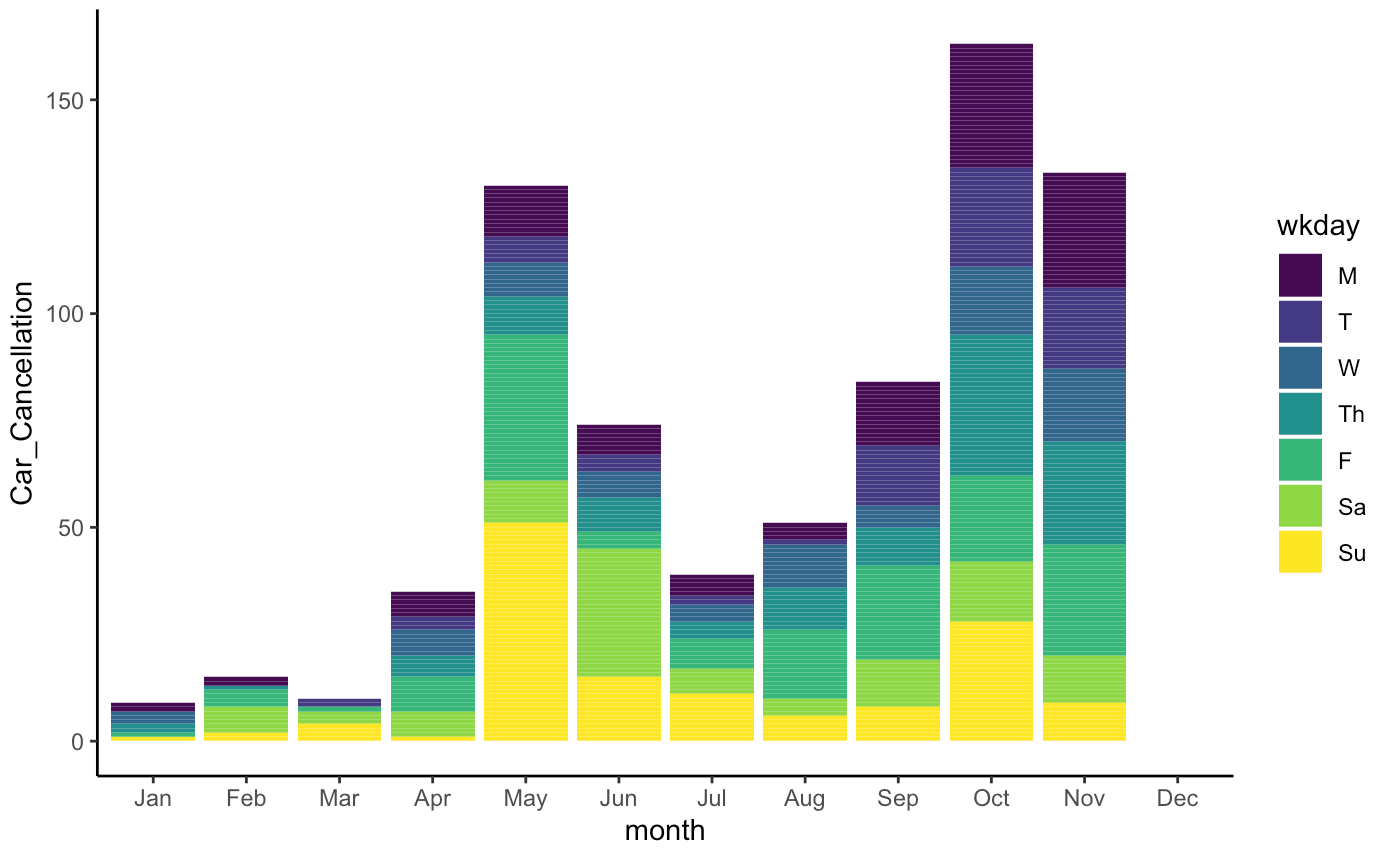


Exhibit 3: Plots for Car\_Cancellation & time (months & weeks)



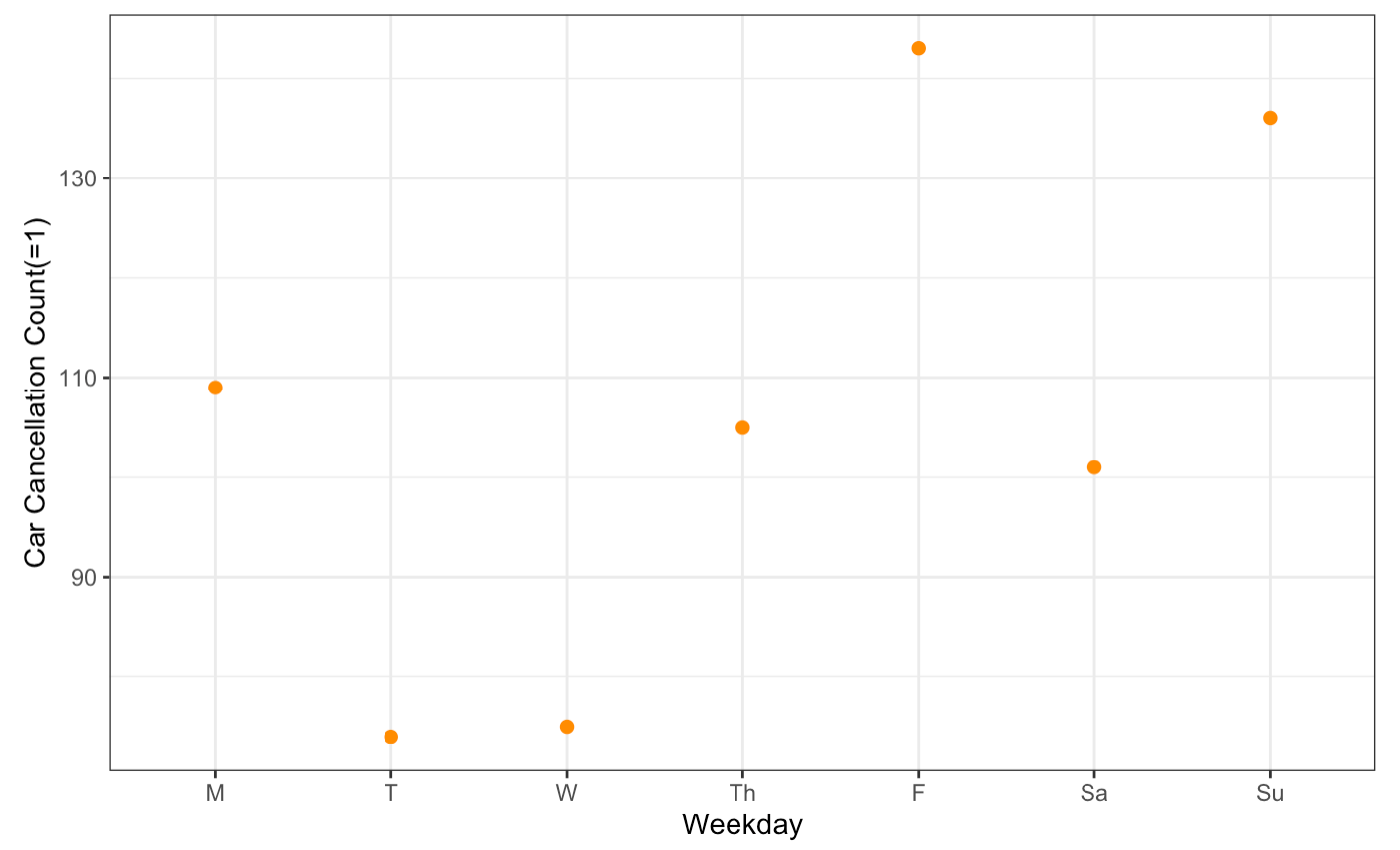
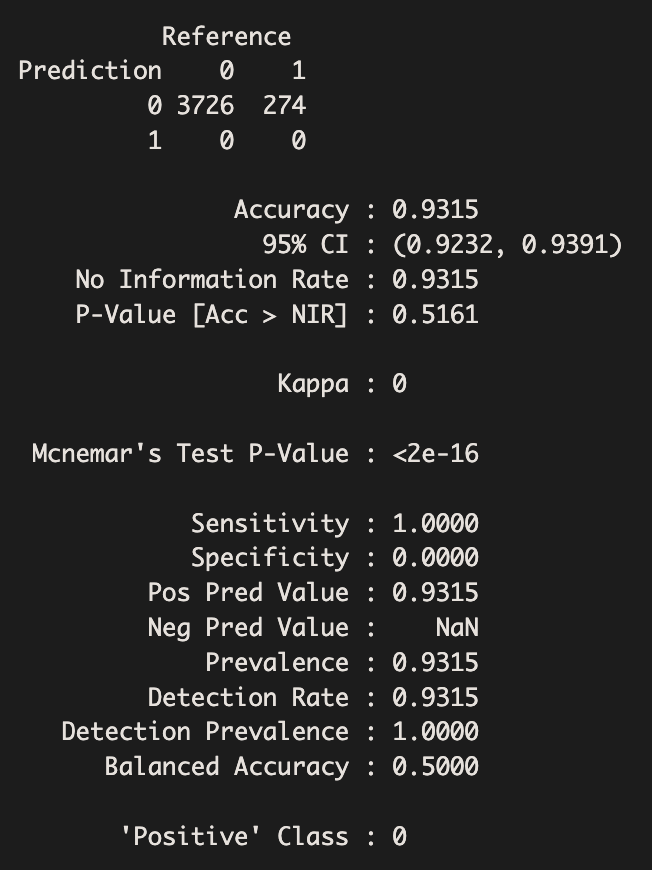
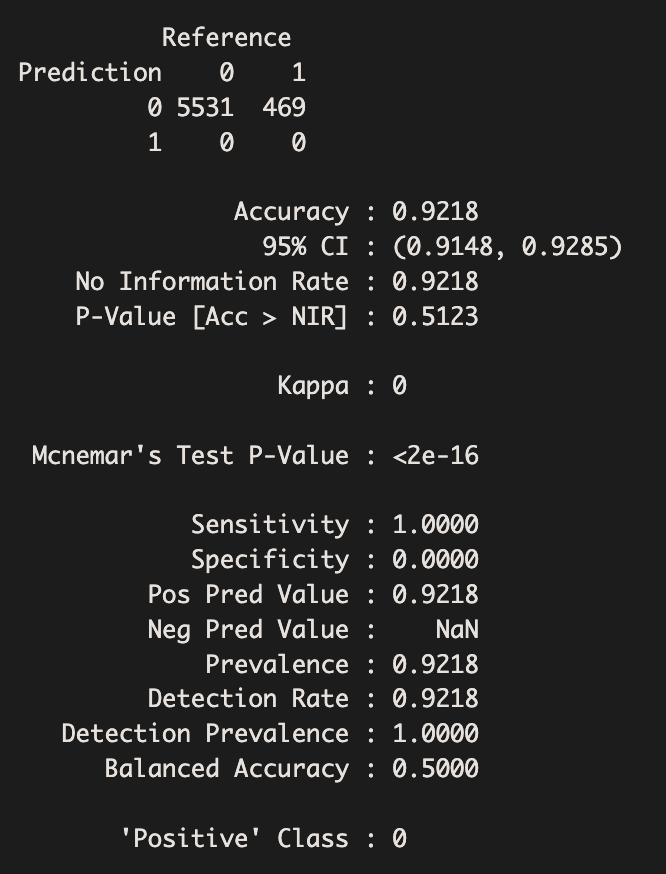
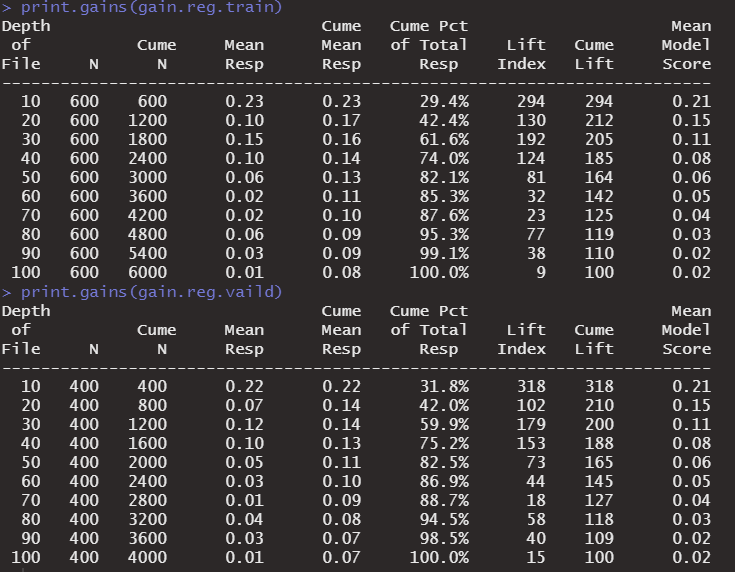
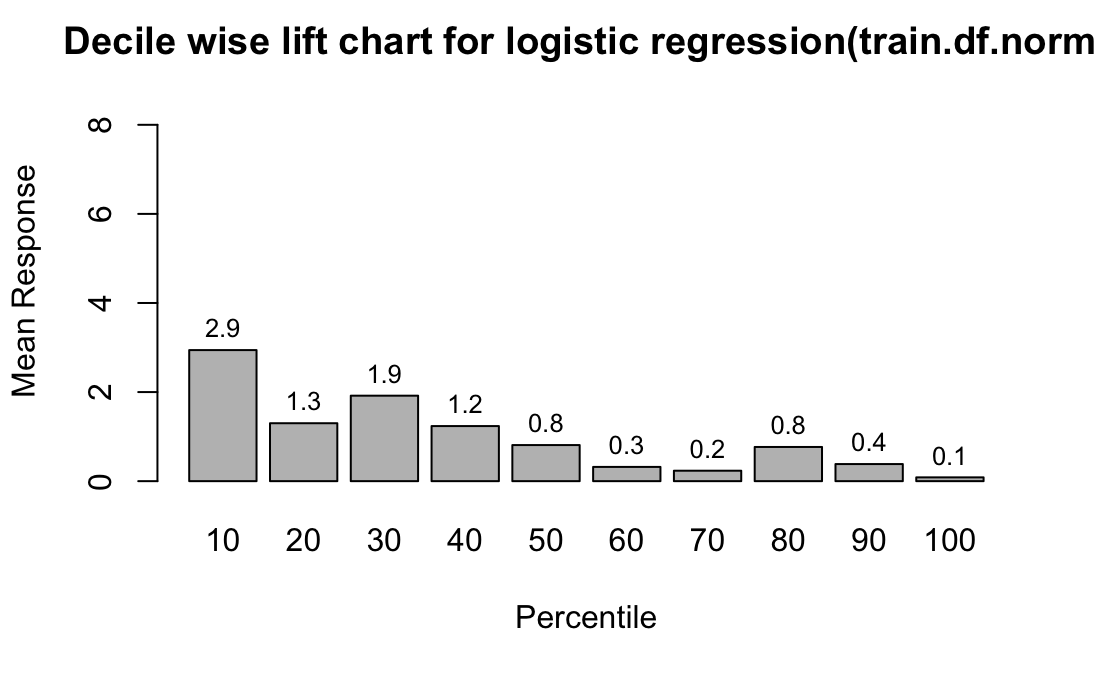
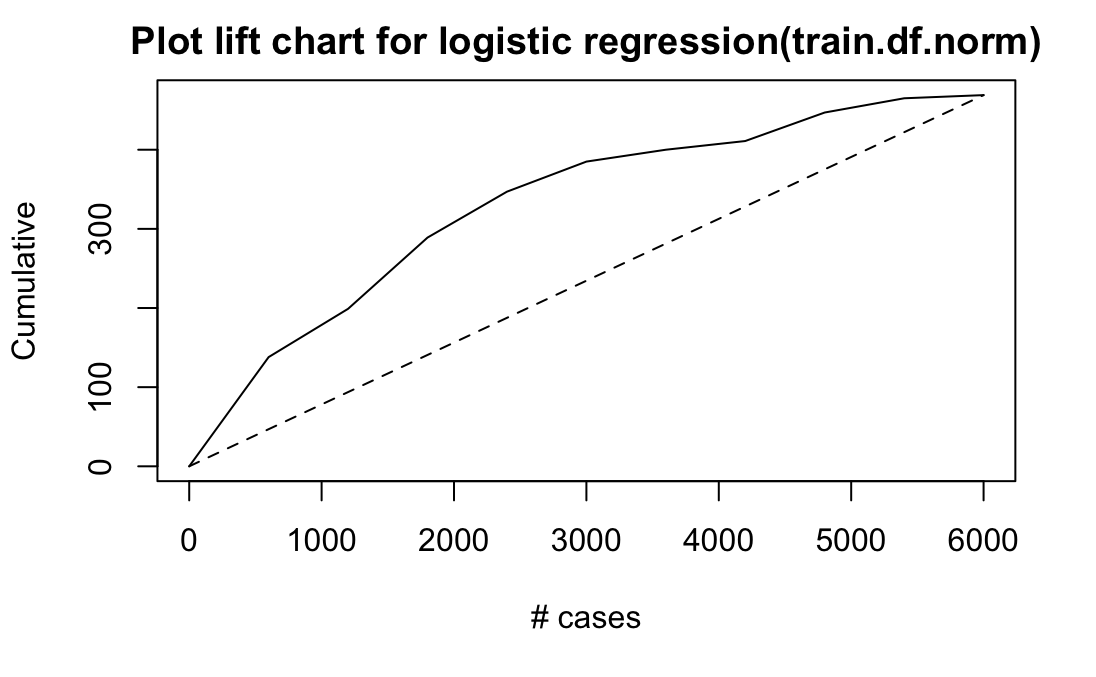


Exhibit 4: Model 1： Logistic Regression Model - Training & Validation



Gain chart for training dataset:



Gain chart for validation dataset:

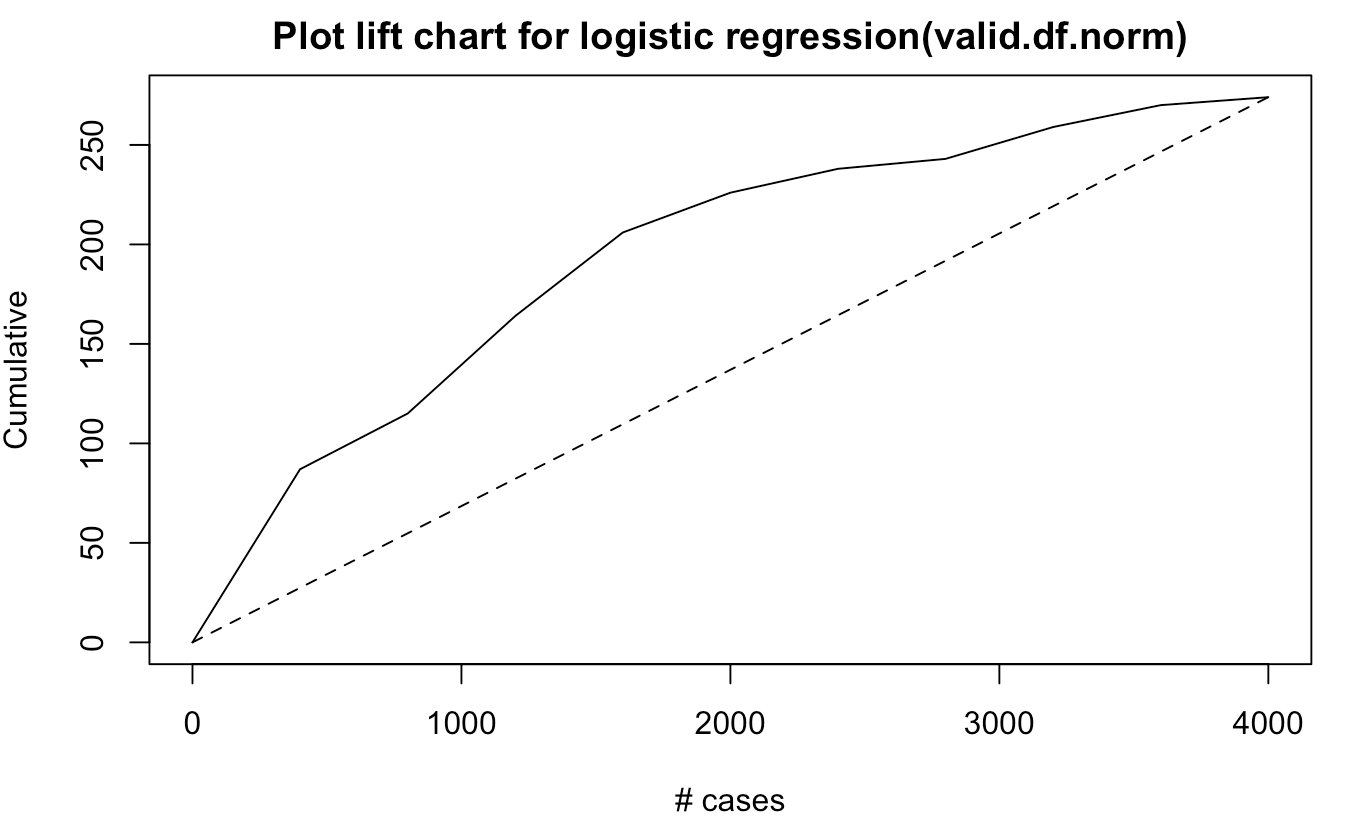
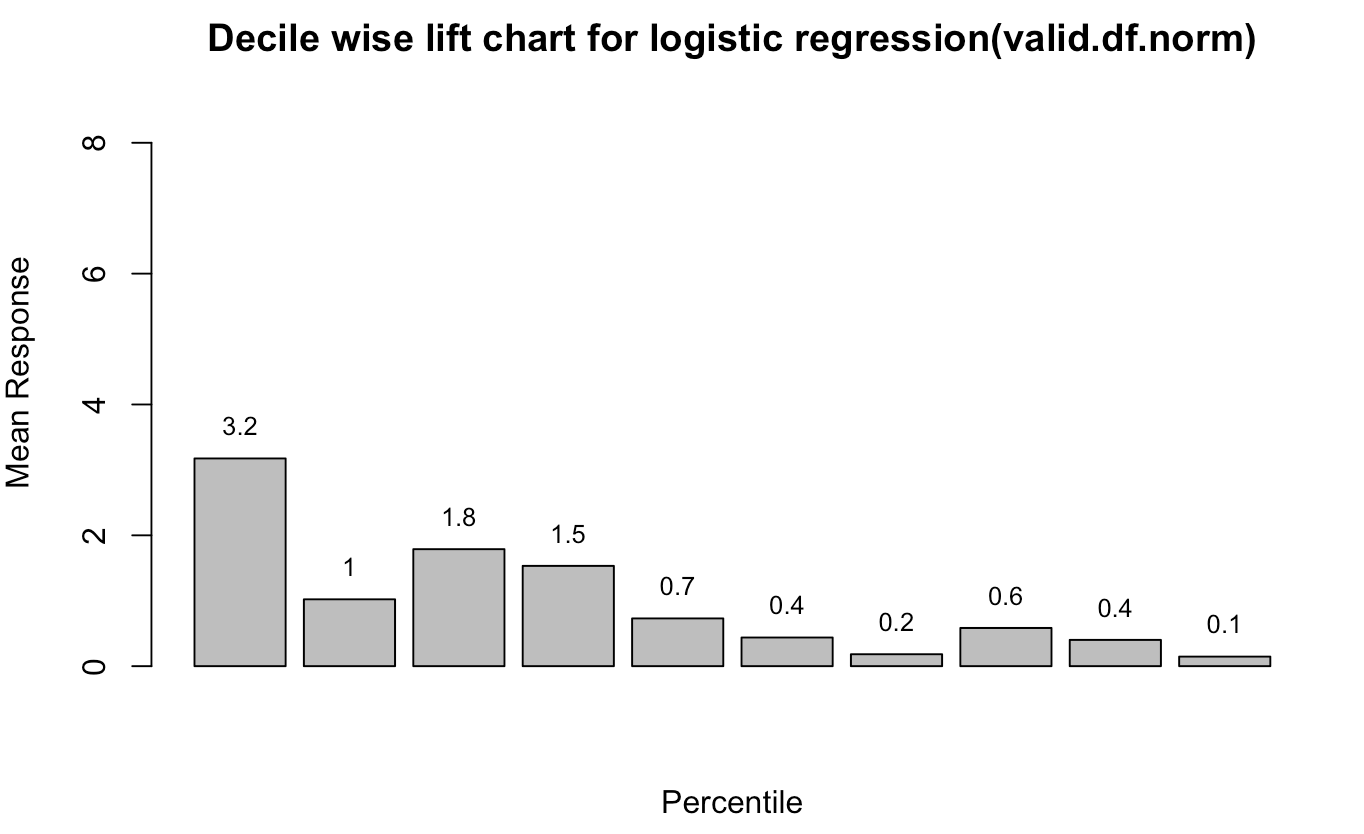
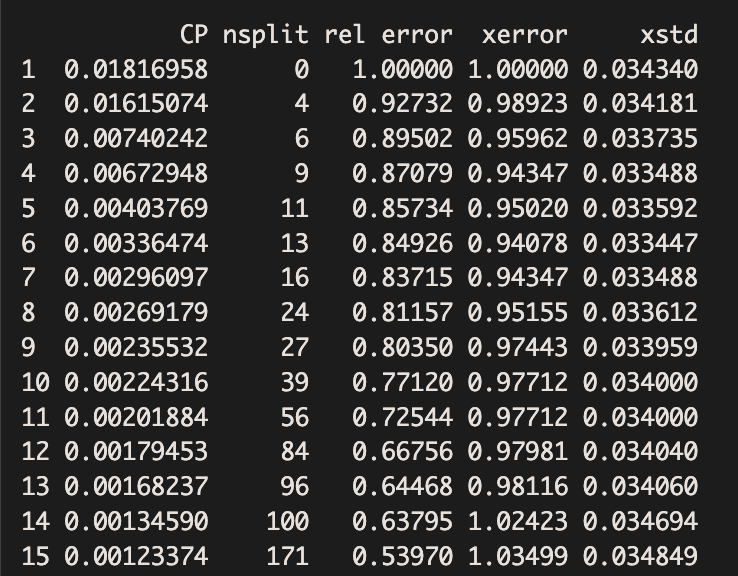
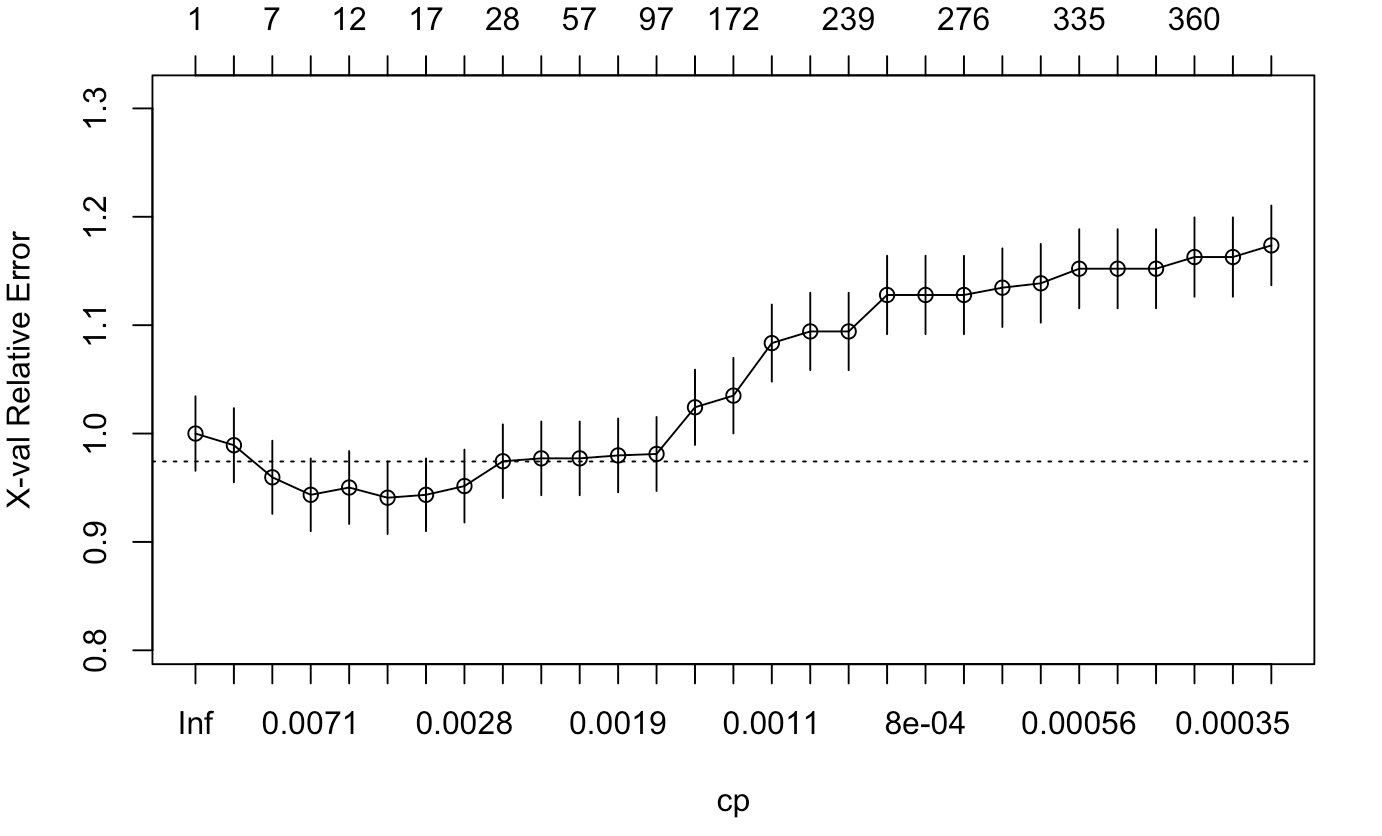


Exhibit 5: Model 2：Classification Tree Model



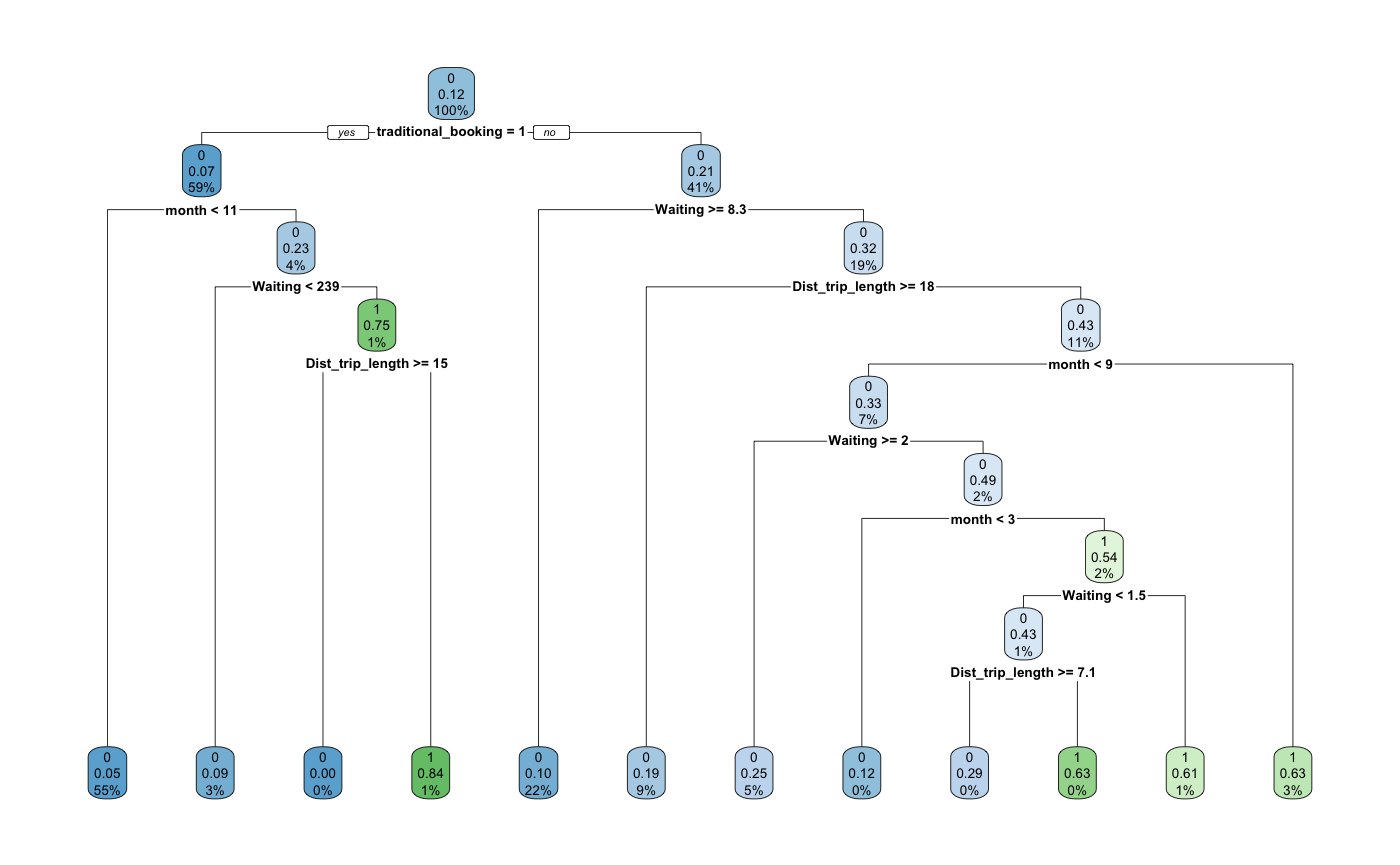
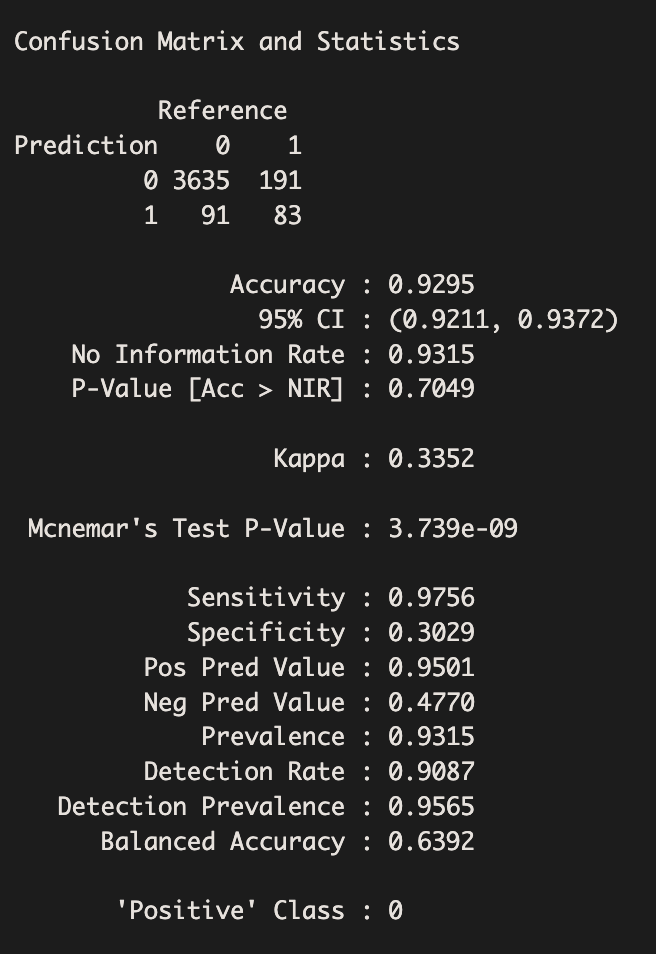
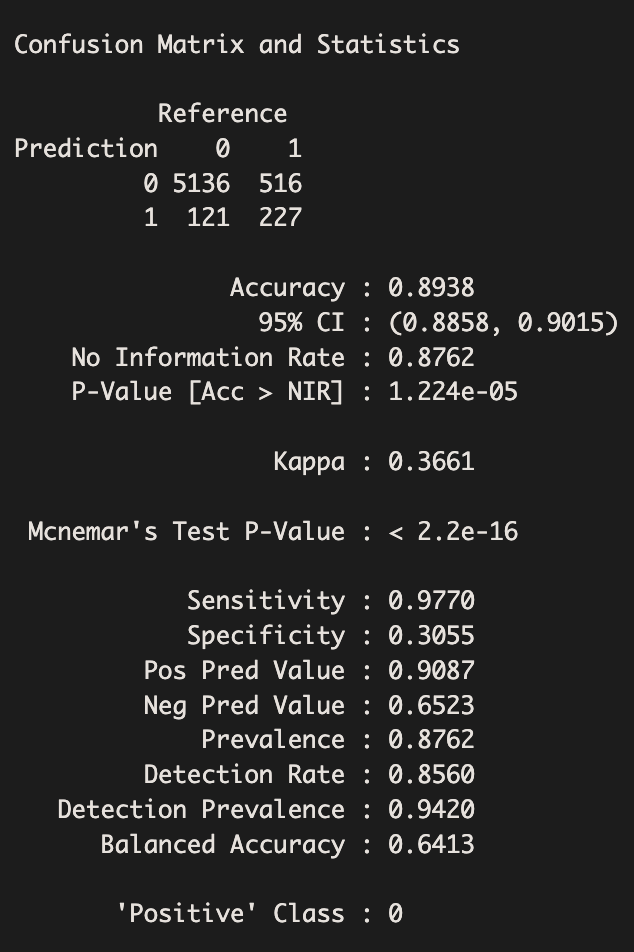
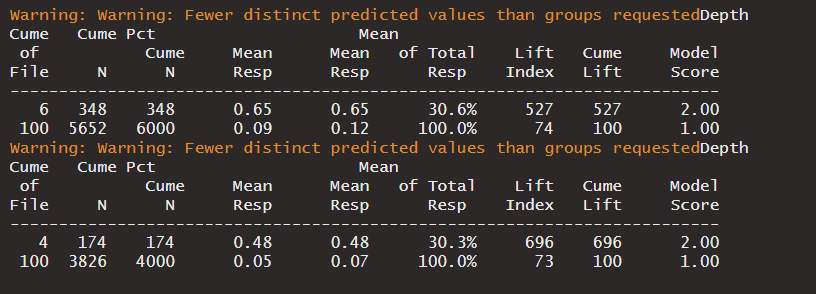
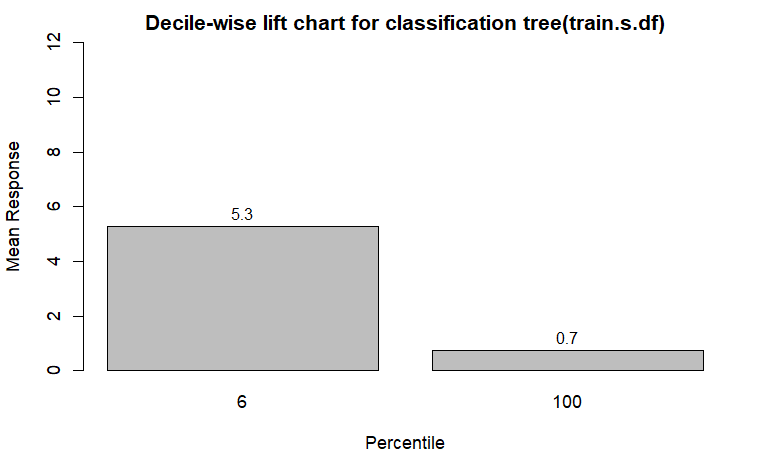
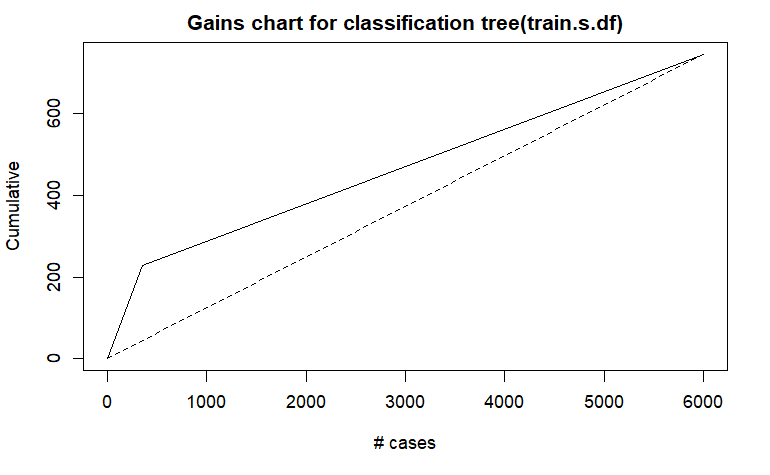
****

Exhibit 6: Model 2：Classification Tree Model for Training & Validation****

****

Gain chart for training dataset:

****

Gain chart for validation dataset:

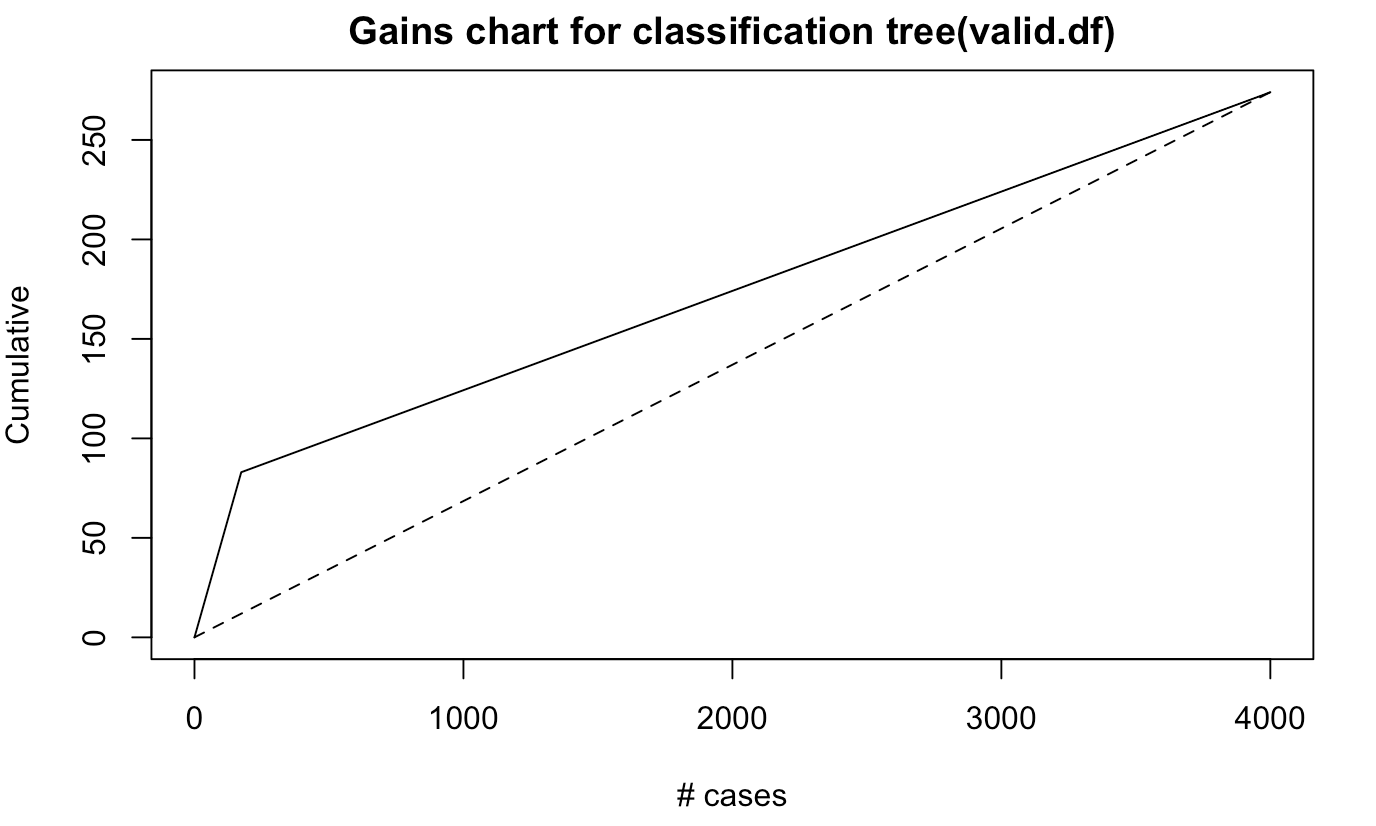
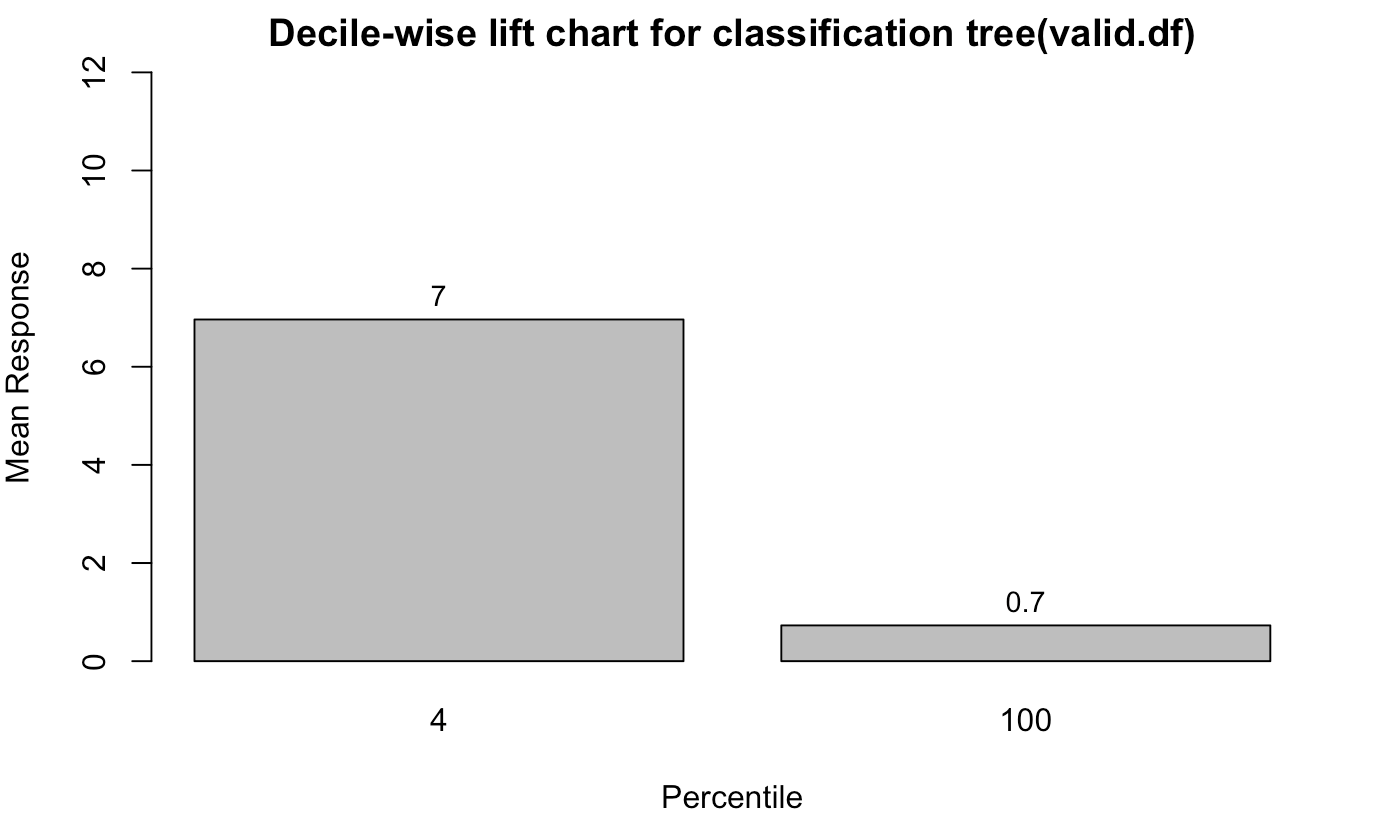
****

Exhibit 7: Model 3: Random Forest Model

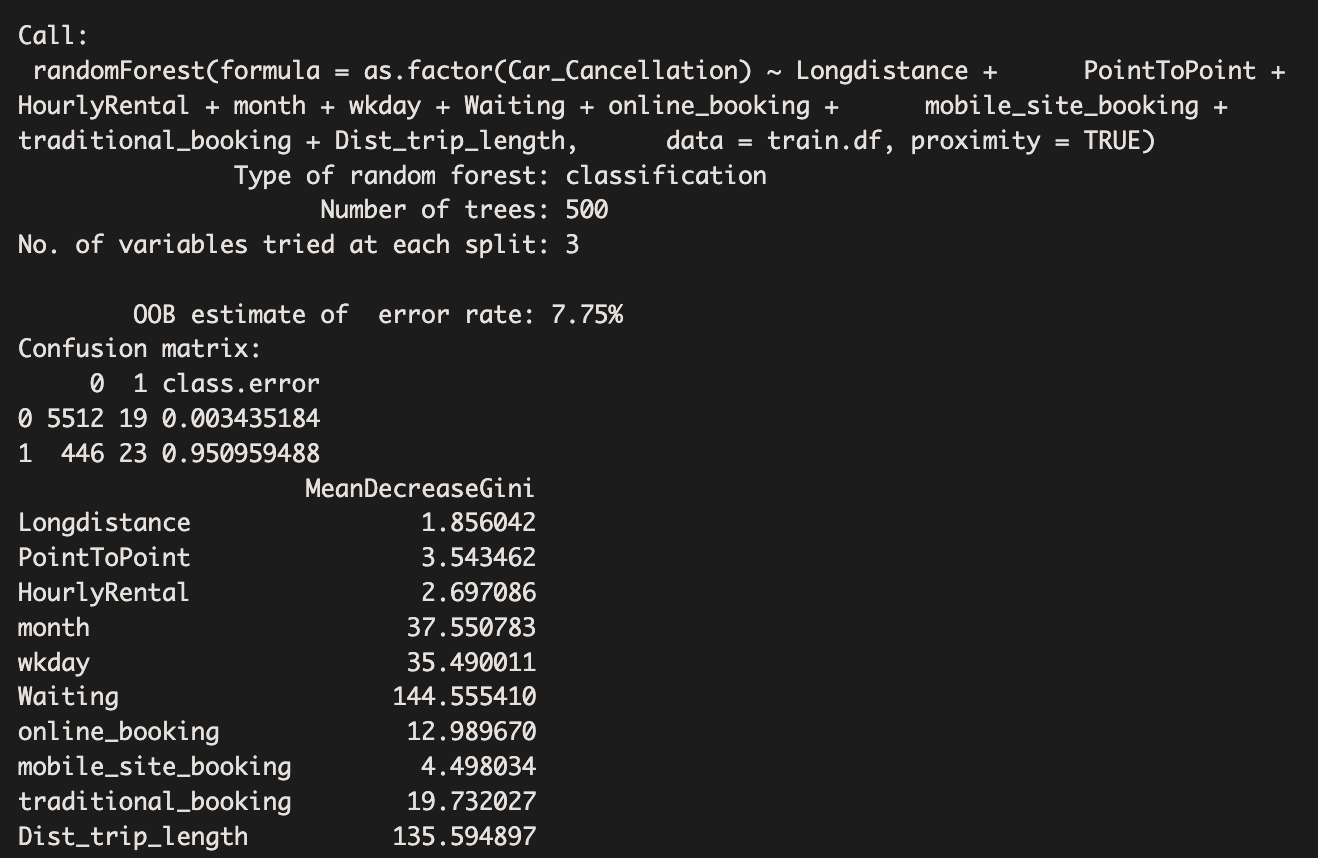
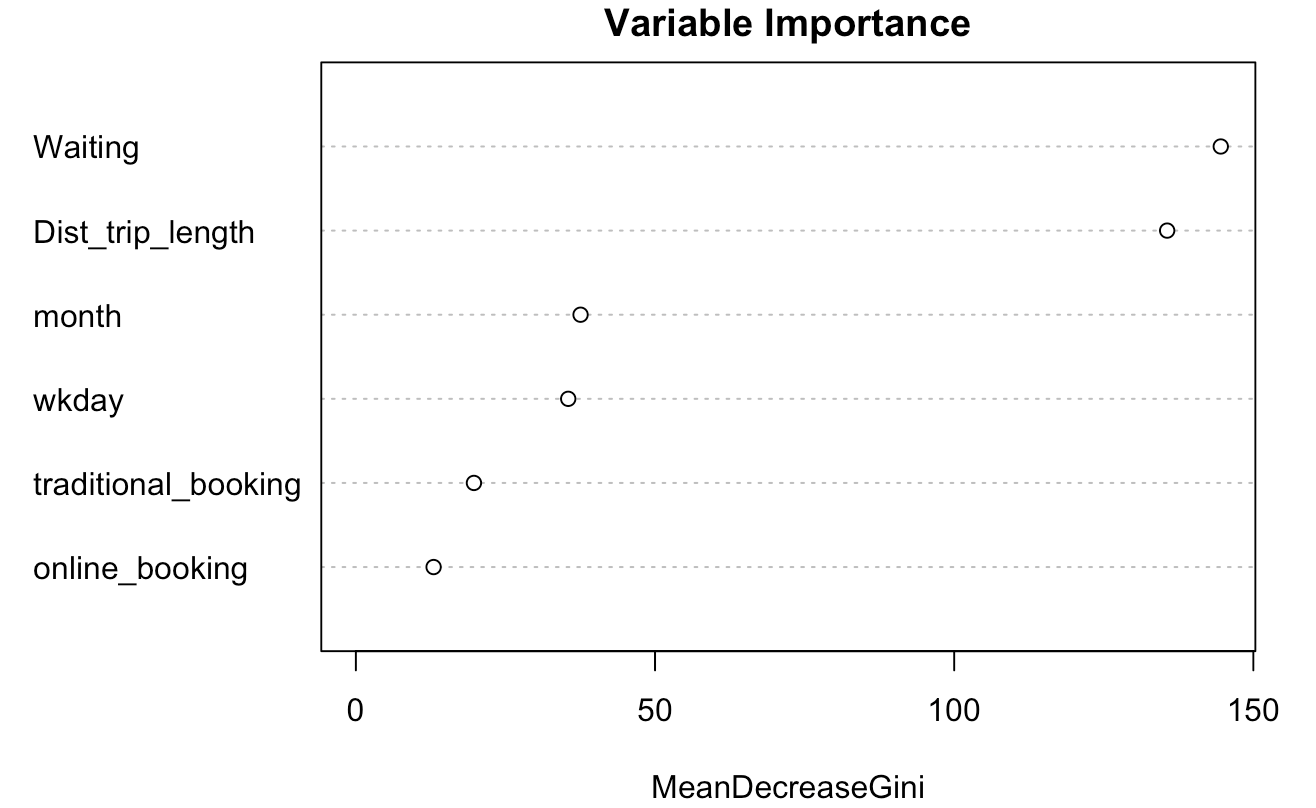


Exhibit 8: Mode 3: Random Forest Model - Confusion matrix for Training & Validation

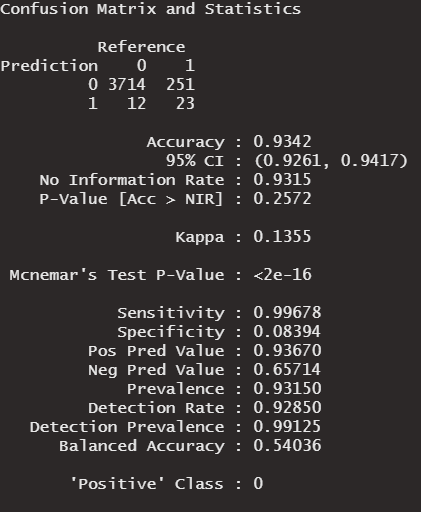
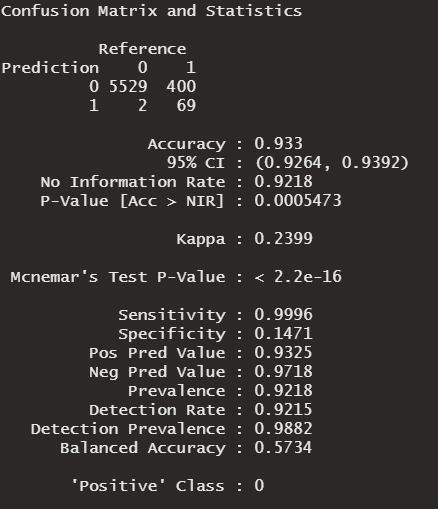
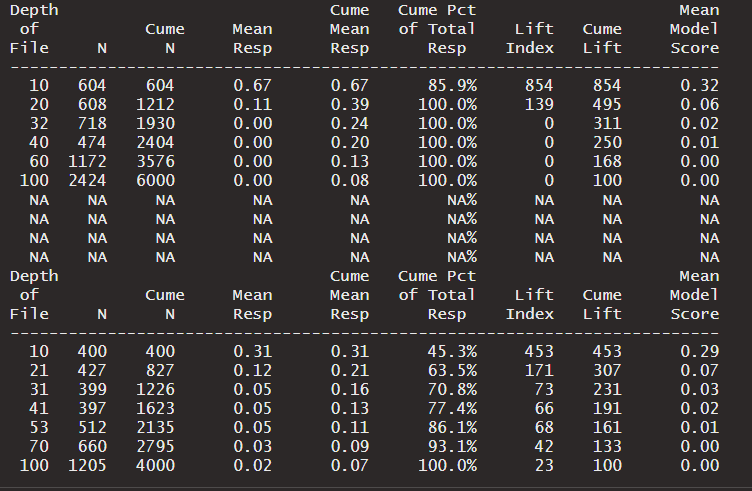
****

Exhibit 9: Model 3: Random Forest - Gain chart and decile chart for Training & Validation 

Gain chart for training dataset:

# 

Gain chart for validation dataset